Preface

How can a manager get out of a data-flooded “mire”? How can a confused decision maker navigate through a “maze”? How can an over-burdened problem solver clean up a “mess”? How can an exhausted scientist bypass a “myth”?

The answer to all of these is to employ a powerful tool known as data mining (DM). DM can turn data into dollars; transform information into intelligence; change patterns into profit; and convert relationships into resources.

As the third branch of operations research and management science (OR/MS) and the third milestone of data management, DM can help support the third category of decision making by elevating raw data into the third stage of knowledge creation.

The term “third” has been mentioned four times above. Let’s go backward and look at three stages of knowledge creation. Managers are drowning in data (the first stage) yet starved for knowledge. A collection of data is not information (the second stage); yet a collection of information is not knowledge! Data are full of information which can yield useful knowledge. The whole subject of DM therefore has a synergy of its own and represents more than the sum of its parts.

There are three categories of decision making: structured, semi-structured and unstructured. Decision making processes fall along a continuum that range from highly structured decisions (sometimes called programmed) to much unstructured, non-programmed decision making (Turban et al., 2006).

At one end of the spectrum, structured processes are routine, often repetitive, problems for which standard solutions exist. Unfortunately, rather than being static, deterministic and simple, the majority of real world problems are dynamic, probabilistic, and complex. Many professional and personal problems can be classified as unstructured, semi-structured, or somewhere in between. In addition to developing normative models (such as linear programming, economic order quantity) for solving structured (or programmed) problems, operation researchers and management scientists have created many descriptive models, such as simulation and goal programming, to deal with semi-structured tasks. Unstructured problems, however, fall into a gray area for which there is no cut-and-dry solution. The current two branches of OR/MS often cannot solve unstructured problems effectively.

To obtain knowledge, one must understand the patterns that emerge from information. Patterns are not just simple relationships among data; they exist separately from information, as archetypes or standards to which emerging information can be compared so that one may draw inferences and take action. Over the last 40 years, the tools and techniques used to process data and information have continued to evolve from databases (DBs) to data warehousing (DW), to DM. DW applications, as a result, have become business-critical and can deliver even more value from these huge repositories of data.

Certainly, there are many statistical models that have emerged over time. Earlier, machine learning has marked a milestone in the evolution of computer science (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996). Although DM is still in its infancy, it is now being used in a wide range of industries and for a range of tasks and contexts (Wang, 2006). DM is synonymous with knowledge discovery in databases, knowledge extraction, data/pattern analysis, data archeology, data dredging, data snooping, data fishing, information harvesting, and
business intelligence (Hand et al., 2001; Giudici, 2003; Han & Kamber, 2006). Data warehousing and mining (DWM) is the science of managing and analyzing large datasets and discovering novel patterns within them. In recent years, DWM has emerged as a particularly exciting and relevant area of research. Prodigious amounts of data are now being generated in domains as diverse and elusive as market research, functional genomics, and pharmaceuticals and intelligently analyzing them to discover knowledge is the challenge that lies ahead.

Yet managing this flood of data, and making it useful and available to decision makers has been a major organizational challenge. We are facing and witnessing global trends (e.g. an information/knowledge-based economy, globalization, technological advances etc.) that drive/motivate data mining and data warehousing research and practice. These developments pose huge challenges (eg. need for faster learning, performance efficiency/effectiveness, new knowledge and innovation) and demonstrate the importance and role of DWM in responding to and aiding this new economy through the use of technology and computing power. DWM allows the extraction of “nuggets” or “pearls” of knowledge from huge historical stores of data. It can help to predict outcomes of future situations, to optimize business decisions, to increase the customer relationship management, and to improve customer satisfaction. As such, DWM has become an indispensable technology for businesses and researchers in many fields.

The Encyclopedia of Data Warehousing and Mining (2nd Edition) provides theories, methodologies, functionalities, and applications to decision makers, problem solvers, and data mining professionals and researchers in business, academia, and government. Since DWM lies at the junction of database systems, artificial intelligence, machine learning and applied statistics, it has the potential to be a highly valuable area for researchers and practitioners. Together with a comprehensive overview, The Encyclopedia of Data Warehousing and Mining (2nd Edition) offers a thorough exposure to the issues of importance in this rapidly changing field. The encyclopedia also includes a rich mix of introductory and advanced topics while providing a comprehensive source of technical, functional, and legal references to DWM.

After spending more than two years preparing this volume, using a totally peer-reviewed process, I am pleased to see it published. Of the 324 articles, there are 214 brand-new articles and 110 updated ones that were chosen from the 234 manuscripts in the first edition. Clearly, the need to significantly update the encyclopedia is due to the tremendous progress in this ever-growing field. Our selection standards were very high. Each chapter was evaluated by at least three peer reviewers; additional third-party reviews were sought in cases of controversy. There have been numerous instances where this feedback has helped to improve the quality of the content, and guided authors on how they should approach their topics. The primary objective of this encyclopedia is to explore the myriad of issues regarding DWM. A broad spectrum of practitioners, managers, scientists, educators, and graduate students who teach, perform research, and/or implement these methods and concepts, can all benefit from this encyclopedia.

The encyclopedia contains a total of 324 articles, written by an international team of 555 experts including leading scientists and talented young scholars from over forty countries. They have contributed great effort to create a source of solid, practical information source, grounded by underlying theories that should become a resource for all people involved in this dynamic new field. Let’s take a peek at a few articles:

Kamel presents an overview of the most important issues and considerations for preparing data for DM. Practical experience of DM has revealed that preparing data is the most time-consuming phase of any DM project. Estimates of the amount of time and resources spent on data preparation vary from at least 60% to upward of 80%. In spite of this fact, not enough attention is given to this important task, thus perpetuating the idea that the core of the DM effort is the modeling process rather than all phases of the DM life cycle.

The past decade has seen a steady increase in the number of fielded applications of predictive DM. The success of such applications depends heavily on the selection and combination of suitable pre-processing and modeling algorithms. Since the expertise necessary for this selection is seldom available in-house, users must either resort to trial-and-error or consultation of experts. Clearly, neither solution is completely satisfactory for the non-expert end-users who wish to access the technology more directly and cost-effectively. Automatic and systematic guidance is required. Giraud-Carrier, Brazdil, Soares, and Vilalta show how meta-learning can be leveraged to provide such guidance through effective exploitation of meta-knowledge acquired through experience.
Ruqian Lu has developed a methodology of acquiring knowledge automatically based on pseudo-natural language understanding. He has won two first class awards from the Academia Sinica and a National second class prize. He has also won the sixth Hua Loo-keng Mathematics Prize.

Wu, McGinnity, and Prasad present a general self-organizing computing network, which have been applied to a hybrid of numerical machine learning approaches and symbolic AI techniques to discover knowledge from databases with a diversity of data types. The authors have also studied various types of bio-inspired intelligent computational models and uncertainty reasoning theories. Based on the research results, the IFOMIND robot control system won the 2005 Fourth British Computer Society’s Annual Prize for Progress towards Machine Intelligence.

Zhang, Xu, and Wang introduce a class of new data distortion techniques based on matrix decomposition. They pioneer use of Singular Value Decomposition and Nonnegative Matrix Factorization techniques for perturbing numerical data values in privacy-preserving DM. The major advantage of this class of data distortion techniques is that they perturb the data as an entire dataset, which is different from commonly used data perturbation techniques in statistics.

There are often situations with large amounts of “unlabeled data” (where only the explanatory variables are known, but the target variable is not known) and with small amounts of labeled data. As recent research in machine learning has shown, using only labeled data to build predictive models can potentially ignore useful information contained in the unlabeled data. Yang and Padmanabhan show how learning patterns from the entire data (labeled plus unlabeled) can be one effective way of exploiting the unlabeled data when building predictive models.

Pratihar explains the principles of some of the non-linear Dimensionality Reduction (DR) techniques, namely Sammon’s Non-Linear Mapping (NLM), VISOR algorithm, Self-Organizing Map (SOM) and Genetic Algorithm (GA)-Like Technique. Their performances have been compared in terms of accuracy in mapping, visibility and computational complexity on a test function – Schaffer’s F1. The author had proposed the above GA-like Technique, previously.

A lot of projected clustering algorithms that focus on finding specific projection for each cluster have been proposed very recently. Deng and Wu found in their study that, besides distance, the closeness of points in different dimensions also depends on the distributions of data along those dimensions. Based on this finding, they propose a projected clustering algorithm, IPROCLUS (Improved PROCLUS), which is efficient and accurate in handling data in high dimensional space. According to the experimental results on real biological data, their algorithm shows much better accuracy than PROCLUS.

Meisel and Mattfeld highlight and summarize the state of the art in attempts to gain synergies from integrating DM and Operations Research. They identify three basic ways of integrating the two paradigms as well as discuss and classify, according to the established framework, recent publications on the intersection of DM and Operations Research.

Yuksektepe and Turkay present a new data classification method based on mixed-integer programming. Traditional approaches that are based on partitioning the data sets into two groups perform poorly for multi-class data classification problems. The proposed approach is based on the use of hyper-boxes for defining boundaries of the classes that include all or some of the points in that set. A mixed-integer programming model is developed for representing existence of hyper-boxes and their boundaries.

Reddy and Rajaratnam give an overview of the Expectation Maximization (EM) algorithm, deriving its theoretical properties, and discussing some of the popularly used global optimization methods in the context of this algorithm. In addition the article provides details of using the EM algorithm in the context of the finite mixture models, as well as a comprehensive set of derivations in the context of Gaussian mixture models. Also, it shows some comparative results on the performance of the EM algorithm when used along with popular global optimization methods for obtaining maximum likelihood estimates and the future research trends in the EM literature.

Smirnov, Pashkin, Levashova, Kashevnik, and Shilov describe usage of an ontology-based context model for decision support purposes and document ongoing research in the area of intelligent decision support based on context-driven knowledge and information integration from distributed sources. Within the research the context is used for representation of a decision situation to the decision maker and for support of the decision maker in
solving tasks typical for the presented situation. The solutions and the final decision are stored in the user profile for further analysis via decision mining to improve the quality of the decision support process.

Corresponding to Feng, XML-enabled association rule framework extends the notion of associated items to XML fragments to present associations among trees rather than simple-structured items of atomic values. They are more flexible and powerful in representing both simple and complex structured association relationships inherent in XML data. Compared with traditional association mining in the well-structured world, mining from XML data, however, is confronted with more challenges due to the inherent flexibilities of XML in both structure and semantics. To make XML-enabled association rule mining truly practical and computationally tractable, template-guided mining of association rules from large XML data must be developed.

With the XML becoming a standard for representing business data, a new trend toward XML DW has been emerging for a couple of years, as well as efforts for extending the XQuery language with near-OLAP capabilities. Mahboubi, Hachicha, and Darmont present an overview of the major XML warehousing approaches, as well as the existing approaches for performing OLAP analyses over XML data. They also discuss the issues and future trends in this area and illustrate this topic by presenting the design of a unified, XML DW architecture and a set of XOLAP operators.

Due to the growing use of XML data for data storage and exchange, there is an imminent need for developing efficient algorithms to perform DM on semi-structured XML data. However, the complexity of its structure makes mining on XML much more complicated than mining on relational data. Ding discusses the problems and challenges in XML DM and provides an overview of various approaches to XML mining.

Pon, Cardenas, and Buttler address the unique challenges and issues involved in personalized online news recommendation, providing background on the shortfalls of existing news recommendation systems, traditional document adaptive filtering, as well as document classification, the need for online feature selection and efficient streaming document classification, and feature extraction algorithms. In light of these challenges, possible machine learning solutions are explored, including how existing techniques can be applied to some of the problems related to online news recommendation.

Clustering is a DM technique to group a set of data objects into classes of similar data objects. While Peer-to-Peer systems have emerged as a new technique for information sharing on Internet, the issues of peer-to-peer clustering have been considered only recently. Li and Lee discuss the main issues of peer-to-peer clustering and reviews representation models and communication models which are important in peer-to-peer clustering.

Users must often refine queries to improve search result relevancy. Query expansion approaches help users with this task by suggesting refinement terms or automatically modifying the user’s query. Finding refinement terms involves mining a diverse range of data including page text, query text, user relevancy judgments, historical queries, and user interaction with the search results. The problem is that existing approaches often reduce relevancy by changing the meaning of the query, especially for the complex ones, which are the most likely to need refinement. Fortunately, the most recent research has begun to address complex queries by using semantic knowledge and Crabtree’s paper provides information about the developments of this new research.

Li addresses web presence and evolution through web log analysis, a significant challenge faced by electronic business and electronic commerce given the rapid growth of the WWW and the intensified competition. Techniques are presented to evolve the web presence and to produce ultimately a predictive model such that the evolution of a given web site can be categorized under its particular context for strategic planning. The analysis of web log data has opened new avenues to assist the web administrators and designers to establish adaptive web presence and evolution to fit user requirements.

It is of great importance to process the raw web log data in an appropriate way, and identify the target information intelligently. Huang, An, and Liu focus on exploiting web log sessions, defined as a group of requests made by a single user for a single navigation purpose, in web usage mining. They also compare some of the state-of-the-art techniques in identifying log sessions from Web servers, and present some applications with various types of Web log data.

Yang has observed that it is hard to organize a website such that pages are located where users expect to find them. Through web usages mining, the authors can automatically discover pages in a website whose location
is different from where users expect to find them. This problem of matching website organization with user expectations is pervasive across most websites.

The Semantic Web technologies provide several solutions concerning the retrieval of Semantic Web Documents (SWDs, mainly, ontologies), which however presuppose that the query is given in a structured way - using a formal language - and provide no advanced means for the (semantic) alignment of the query to the contents of the SWDs. Kotis reports on recent research towards supporting users to form semantic queries – requiring no knowledge and skills for expressing queries in a formal language - and to retrieve SWDs whose content is similar to the queries formed.

Zhu, Nie, and Zhang noticed that extracting object information from the Web is of significant importance. However, the diversity and lack of grammars of Web data make this task very challenging. Statistical Web object extraction is a framework based on statistical machine learning theory. The potential advantages of statistical Web object extraction models lie in the fact that Web data have plenty of structure information and the attributes about an object have statistically significant dependencies. These dependencies can be effectively incorporated by developing an appropriate graphical model and thus result in highly accurate extractors.

Borges and Levene advocate the use of Variable Length Markov Chains (VLMC) models for Web usage mining since they provide a compact and powerful platform for Web usage analysis. The authors review recent research methods that build VLMC models, as well as methods devised to evaluate both the prediction power and the summarization ability of a VLMC model induced from a collection of navigation sessions. Borges and Levene suggest that due to the well established concepts from Markov chain theory that underpin VLMC models, they will be capable of providing support to cope with the new challenges in Web mining.

With the rapid growth of online information (i.e. web sites, textual document) text categorization has become one of the key techniques for handling and organizing data in textual format. First fundamental step in every activity of text analysis is to transform the original file in a classical database, keeping the single words as variables. Cerchiello presents the current state of the art, taking into account all the available classification methods and offering some hints on the more recent approaches. Also, Song discusses issues and methods in automatic text categorization, which is the automatic assigning of pre-existing category labels to a group of documents. The article reviews the major models in the field, such as naïve Bayesian classifiers, decision rule classifiers, the $k$-nearest neighbor algorithm, and support vector machines. It also outlines the steps requires to prepare a text classifier and touches on related issues such as dimensionality reduction and machine learning techniques.

Sentiment analysis refers to the classification of texts based on the sentiments they contain. It is an emerging research area in text mining and computational linguistics, and has attracted considerable research attention in the past few years. Leung and Chan introduce a typical sentiment analysis model consisting of three core steps, namely data preparation, review analysis and sentiment classification, and describes representative techniques involved in those steps.

Yu, Tungare, Fan, Pérez- Quiñones, Fox, Cameron, and Cassel describe text classification on a specific information genre, one of the text mining technologies, which is useful in genre-specific information search. Their particular interest is on course syllabus genre. They hope their work is helpful for other genre-specific classification tasks.

Hierarchical models have been shown to be effective in content classification. However, an empirical study has shown that the performance of a hierarchical model varies with given taxonomies; even a semantically sound taxonomy has potential to change its structure for better classification. Tang and Liu elucidate why a given semantics-based hierarchy may not work well in content classification, and how it could be improved for accurate hierarchical classification.

Serrano and Castillo present a survey on the most recent methods to index documents written in natural language to be dealt by text mining algorithms. Although these new indexing methods, mainly based of hyperspaces of word semantic relationships, are a clear improvement on the traditional “bag of words” text representation, they are still producing representations far away from the human mind structures. Future text indexing methods should take more aspects from human mind procedures to gain a higher level of abstraction and semantic depth to success in free-text mining tasks.
Pan presents recent advances in applying machine learning and DM approaches to extract automatically explicit and implicit temporal information from natural language text. The extracted temporal information includes, for example, events, temporal expressions, temporal relations, (vague) event durations, event anchoring, and event orderings.

Saxena, Kothari, and Pandey present a brief survey of various techniques that have been used in the area of Dimensionality Reduction (DR). In it, evolutionary computing approach in general, and Genetic Algorithm in particular have been used as approach to achieve DR.

Huang, Krneta, Lin, and Wu describe the notion of Association Bundle Identification. Association bundles were presented by Huang et al. (2006) as a new pattern of association for DM. On applications such as the Market Basket Analysis, association bundles can be compared to, but essentially distinguished from the well-established association rules. Association bundles present meaningful and important associations that association rules unable to identify.

Bartík and Zendulka analyze the problem of association rule mining in relational tables. Discretization of quantitative attributes is a crucial step of this process. Existing discretization methods are summarized. Then, a method called Average Distance Based Method, which was developed by the authors, is described in detail. The basic idea of the new method is to separate processing of categorical and quantitative attributes. A new measure called average distance is used during the discretization process.

Leung provides a comprehensive overview of constraint-based association rule mining, which aims to find interesting relationships—represented by association rules that satisfy user-specified constraints—among items in a database of transactions. The author describes what types of constraints can be specified by users and discusses how the properties of these constraints can be exploited for efficient mining of interesting rules.

Pattern discovery was established for second order event associations in early 90’s by the authors’ research group (Wong, Wang, and Li). A higher order pattern discovery algorithm was devised in the mid 90s for discrete-valued data sets. The discovered high order patterns can then be used for classification. The methodology was later extended to continuous and mixed-mode data. Pattern discovery has been applied in numerous real-world and commercial applications and is an ideal tool to uncover subtle and useful patterns in a database.

Li and Ng discuss the Positive Unlabelled learning problem. In practice, it is costly to obtain the class labels for large sets of training examples, and oftentimes the negative examples are lacking. Such practical considerations motivate the development of a new set of classification algorithms that can learn from a set of labeled positive examples P augmented with a set of unlabeled examples U. Four different techniques, S-EM, PEBL, Roc-SVM and LPLP, have been presented. Particularly, LPLP method was designed to address a real-world classification application where the size of positive examples is small.

The classification methodology proposed by Yen aims at using different similarity information matrices extracted from citation, author, and term frequency analysis for scientific literature. These similarity matrices were fused into one generalized similarity matrix by using parameters obtained from a genetic search. The final similarity matrix was passed to an agglomerative hierarchical clustering routine to classify the articles. The work, synergistically integrates multiple similarity information, showed that the proposed method was able to identify the main research disciplines, emerging fields, major contributing authors and their area of expertise within the scientific literature collection.

As computationally intensive experiments are increasingly found to incorporate massive data from multiple sources, the handling of original data, the derived data and all intermediate datasets became challenging. Data provenance is a special kind of Metadata that holds information about who did what and when. Sorathia and Maitra discuss various methods, protocols and system architecture for data provenance. It provides insights about how data provenance can affect decisions for utilization. From recent research perspective, it introduces how grid based data provenance can provide effective solution for data provenance even in Service Orientation Paradigm.

The practical usages of Frequent Pattern Mining (FPM) algorithms in knowledge mining tasks are still limited due to the lack of interpretability caused from the enormous output size. Conversely, we observed recently a growth of interest in FPM community to summarize the output of an FPM algorithm and obtain a smaller set
of patterns that is non-redundant, discriminative, and representative (of the entire pattern set). Hasan surveys
different summarization techniques with a comparative discussion among their benefits and limitations.

Data streams are usually generated in an online fashion characterized by huge volume, rapid unpredictable
rates, and fast changing data characteristics. Dang, Ng, Ong, and Lee discuss this challenge in the context of
finding frequent sets from transactional data streams. In it, some effective methods are reviewed and discussed,
in three fundamental mining models for data stream environments: landmark window, forgetful window and
sliding window models.

Research in association rules mining has initially concentrated in solving the obvious problem of finding
positive association rules; that is, rules among items that remain in the transactions. It was only several
years after that the possibility of finding also negative association rules was investigated, based though on the absence
of items from transactions. Ioannis gives an overview of the works having engaged with the subject until now
and present a novel view for the definition of negative influence among items, where the choice of one item can
trigger the removal of another one.

Lin and Tseng consider mining generalized association rules in an evolving environment. They survey dif-
fferent strategies incorporating the state-of-the-art techniques in dealing with this problem and investigate how
to update efficiently the discovered association rules when there is transaction update to the database along with
item taxonomy evolution and refinement of support constraint.

Feature extraction/selection has received considerable attention in various areas for which thousands of features
are available. The main objective of feature extraction/selection is to identify a subset of feature that are most
predictive or informative of a given response variable. Successful implementation of feature extraction/selec-
tion not only provides important information for prediction or classification, but also reduces computational and
analytical efforts for the analysis of high-dimensional data. Kim presents various feature extraction/selection
methods, along with some real examples.

Feature interaction presents a challenge to feature selection for classification. A feature by itself may have
little correlation with the target concept, but when it is combined with some other features; they can be strongly
correlated with the target concept. Unintentional removal of these features can result in poor classification
performance. Handling feature interaction could be computationally intractable. Zhao and Liu provide a com-
prehensive study for the concept of feature interaction and present several existing feature selection algorithms
that apply feature interaction.

François addresses the problem of feature selection in the context of modeling the relationship between ex-
planatory variables and target values, which must be predicted. It introduces some tools, general methodology
to be applied on it and identifies trends and future challenges.

Datasets comprising of many features can lead to serious problems, like low classification accuracy. To ad-
dress such problems, feature selection is used to select a small subset of the most relevant features. The most
widely used feature selection approach is the wrapper, which seeks relevant features by employing a classifier
in the selection process. Chrysostomou, Lee, Chen, and Liu present the state of the art of the wrapper feature
selection process and provide an up-to-date review of work addressing the limitations of the wrapper and im-
proving its performance.

Lisi considers the task of mining multiple-level association rules extended to the more complex case of hav-
ing an ontology as prior knowledge. This novel problem formulation requires algorithms able to deal actually
with ontologies, i.e. without disregarding their nature of logical theories equipped with a formal semantics. Lisi
describes an approach that resorts to the methodological apparatus of that logic-based machine learning form
known under the name of Inductive Logic Programming, and to the expressive power of those knowledge rep-
resentation frameworks that combine logical formalisms for databases and ontologies.

Arslan presents a unifying view for many sequence alignment algorithms in the literature proposed to guide
the alignment process. Guiding finds its true meaning in constrained sequence alignment problems, where
constraints require inclusion of known sequence motifs. Arslan summarizes how constraints have evolved from
inclusion of simple subsequence motifs to inclusion of subsequences within a tolerance, then to more general
regular expression-described motif inclusion, and to inclusion of motifs described by context free grammars.
Xiong, Wang, and Zhang introduce a novel technique to alignment manifolds so as to learn the correspondence relationship in data. The authors argue that it will be more advantageous if they can guide the alignment by relative comparison, which is well defined frequently and easy to obtain. The authors show how this problem can be formulated as an optimization procedure. To make the solution tractable, they further re-formulated it as a convex semi-definite programming problem.

Time series data are typically generated by measuring and monitoring applications and plays a central role in predicting the future behavior of systems. Since time series data in its raw form contain no usable structure, it is often segmented to generate a high-level data representation that can be used for prediction. Chundi, and Rosenkrantz discuss the segmentation problem and outline the current state-of-the-art in generating segmentations for the given time series data.

Customer segmentation is the process of dividing customers into distinct subsets (segments or clusters) that behave in the same way or have similar needs. There may exist natural behavioral patterns in different groups of customers or customer transactions. Yang discusses research on using behavioral patterns to segment customers.

Along the lines of Wright and Stashuk, quantization based schemes seemingly discard important data by grouping individual values into relatively large aggregate groups; the use of fuzzy and rough set tools helps to recover a significant portion of the data lost by performing such a grouping. If quantization is to be used as the underlying method of projecting continuous data into a form usable by a discrete-valued knowledge discovery system, it is always useful to evaluate the benefits provided by including a representation of the vagueness derived from the process of constructing the quantization bins.

Lin provides a comprehensive coverage for one of the important problems in DM: sequential pattern mining, especially in the aspect of time constraints. It gives an introduction to the problem, defines the constraints, reviews the important algorithms for the research issue and discusses future trends.

Chen explores the subject of clustering time series, concentrating specially on the area of subsequence time series clustering; dealing with the surprising recent result that the traditional method used in this area is meaningless. He reviews the results that led to this startling conclusion, reviews subsequent work in the literature dealing with this topic, and goes on to argue that two of these works together form a solution to the dilemma.

Qiu and Malthouse summarize the recent developments in cluster analysis for categorical data. The traditional latent class analysis assumes that manifest variables are independent conditional on the cluster identity. This assumption is often violated in practice. Recent developments in latent class analysis relax this assumption by allowing for flexible correlation structure for manifest variables within each cluster. Applications to real datasets provide easily interpretable results.

Learning with Partial Supervision (LPS) aims at combining labeled and unlabeled data to boost the accuracy of classification and clustering systems. The relevance of LPS is highly appealing in applications where only a small ratio of labeled data and a large number of unlabeled data are available. LPS strives to take advantage of traditional clustering and classification machineries to deal with labeled data scarcity. Bouchachia introduces LPS and outlines the different assumptions and existing methodologies concerning it.

Wei, Li, and Li introduce a novel learning paradigm called enclosing machine learning for DM. The new learning paradigm is motivated by two cognition principles of human being, which are cognizing things of the same kind and, recognizing and accepting things of a new kind easily. The authors made a remarkable contribution setting up a bridge that connects the cognition process understanding, with mathematical machine learning tools under the function equivalence framework.

Bouguettaya and Yu focus on investigating the behavior of agglomerative hierarchical algorithms. They further divide these algorithms into two major categories: group based and single-object based clustering methods. The authors choose UPGMA and SLINK as the representatives of each category and the comparison of these two representative techniques could also reflect some similarity and difference between these two sets of clustering methods. Experiment results show a surprisingly high level of similarity between the two clustering techniques under most combinations of parameter settings.
In an effort to achieve improved classifier accuracy, extensive research has been conducted in classifier ensembles. Cluster ensembles offer a solution to challenges inherent to clustering arising from its ill-posed nature. Domeniconi and Razgan discuss recent developments in ensemble methods for clustering.

Tsoumakas and Vlahavas introduce the research area of Distributed DM (DDM). They present the state-of-the-art DDM methods for classification, regression, association rule mining and clustering and discuss the application of DDM methods in modern distributed computing environments such as the Grid, peer-to-peer networks and sensor networks.

Wu, Xiong, and Chen highlight the relationship between the clustering algorithms and the distribution of the “true” cluster sizes of the data. They demonstrate that k-means tends to show the uniform effect on clusters, whereas UPGMA tends to take the dispersion effect. This study is crucial for the appropriate choice of the clustering schemes in DM practices.

Huang describes k-modes, a popular DM algorithm for clustering categorical data, which is an extension to k-means with modifications on the distance function, representation of cluster centers and the method to update the cluster centers in the iterative clustering process. Similar to k-means, the k-modes algorithm is easy to use and efficient in clustering large data sets. Other variants are also introduced, including the fuzzy k-modes for fuzzy cluster analysis of categorical data, k-prototypes for clustering mixed data with both numeric and categorical values, and W-k-means for automatically weighting attributes in k-means clustering.

Xiong, Steinbach, Tan, Kumar, and Zhou describe a pattern preserving clustering method, which produces interpretable and usable clusters. Indeed, while there are strong patterns in the data---patterns that may be a key for the analysis and description of the data---these patterns are often split among different clusters by current clustering approaches, since clustering algorithms have no built in knowledge of these patterns and may often have goals that are in conflict with preserving patterns. To that end, their focus is to characterize (1) the benefits of pattern preserving clustering and (2) the most effective way of performing pattern preserving clustering.

Semi-supervised clustering uses the limited background knowledge to aid unsupervised clustering algorithms. Recently, a kernel method for semi-supervised clustering has been introduced. However, the setting of the kernel’s parameter is left to manual tuning, and the chosen value can largely affect the quality of the results. Yan and Domeniconi derive a new optimization criterion to automatically determine the optimal parameter of an RBF kernel, directly from the data and the given constraints. The proposed approach integrates the constraints into the clustering objective function, and optimizes the parameter of a Gaussian kernel iteratively during the clustering process.

Vilalta and Stepinski propose a new approach to external cluster validation based on modeling each cluster and class as a probabilistic distribution. The degree of separation between both distributions can then be measured using an information-theoretic approach (e.g., relative entropy or Kullback-Leibler distance). By looking at each cluster individually, one can assess the degree of novelty (large separation to other classes) of each cluster, or instead the degree of validation (close resemblance to other classes) provided by the same cluster.

Casado, Pacheco, and Nuñez have designed a new technique based on the metaheuristic strategy Tabu Search for variable selection for classification, in particular for discriminant analysis and logistic regression. There are very few key references on the selection of variables for their use in discriminant analysis and logistic regression. For this specific purpose only the Stepwise, Backward and Forward methods, can be found in the literature. These methods are simple and they are not very efficient when there are many original variables.

Ensemble learning is an important method of deploying more than one learning model to give improved predictive accuracy for a given learning problem. Rooney, Patterson, and Nugent describe how regression based ensembles are able to reduce the bias and/or variance of the generalization error and review the main techniques that have been developed for the generation and integration of regression based ensembles.

Dominik, Walczak, and Wojciechowski evaluate performance of the most popular and effective classifiers with graph structures, on two kinds of classification problems from different fields of science: computational chemistry, chemical informatics (chemical compounds classification) and information science (web documents classification).
Tong, Koren, and Faloutsos study asymmetric proximity measures on directed graphs, which quantify the relationships between two nodes. Their proximity measure is based on the concept of escape probability. This way, the authors strive to summarize the multiple facets of nodes-proximity, while avoiding some of the pitfalls to which alternative proximity measures are susceptible. A unique feature of the measures is accounting for the underlying directional information. The authors put a special emphasis on computational efficiency, and develop fast solutions that are applicable in several settings and they show the usefulness of their proposed direction-aware proximity method for several applications.

Classification models and in particular binary classification models are ubiquitous in many branches of science and business. Model performance assessment is traditionally accomplishing by using metrics, derived from the confusion matrix or contingency table. It has been observed recently that **Receiver Operating Characteristic (ROC)** curves visually convey the same information as the confusion matrix in much more intuitive and robust fashion. Hamel illustrates how ROC curves can be deployed for model assessment to provide a much deeper and perhaps more intuitive analysis of classification models.

Molecular classification involves the classification of samples into groups of biological phenotypes based on data obtained from microarray experiments. The high-dimensional and multiclass nature of the classification problem demands work on two specific areas: (1) feature selection (FS) and (2) decomposition paradigms. Ooi introduces a concept called **differential prioritization**, which ensures that the optimal balance between two FS criteria, relevance and redundancy, is achieved based on the number of classes in the classification problem.

Incremental learning is a learning strategy that aims at equipping learning systems with adaptively, which allows them to adjust themselves to new environmental conditions. Usually, it implicitly conveys an indication to future evolution and eventually self correction over time as new events happen, new input becomes available, or new operational conditions occur. Bouchachia brings in incremental learning, discusses the main trends of this subject and outlines some of the contributions of the author.

Sheng and Ling introduce the theory of the cost-sensitive learning. The theory focuses on the most common cost (i.e. misclassification cost), which plays the essential role in cost-sensitive learning. Without loss of generality, the authors assume binary classification in this article. Based on the binary classification, they infer that the original cost matrix in real-world applications can always be converted to a simpler one with only false positive and false negative costs.

Thomopoulos focuses on the cooperation of heterogeneous knowledge for the construction of a domain expertise. A two-stage method is proposed: First, verifying expert knowledge (expressed in the conceptual graph model) by experimental data (in the relational model) and second, discovering unexpected knowledge to refine the expertise. A case study has been carried out to further explain the use of this method.

Recupero discusses the graph matching problem and related filtering techniques. It introduces GrepVS, a new fast graph matching algorithm, which combines filtering ideas from other well-known methods in literature. The chapter presents details on hash tables and the Berkeley DB, used to store efficiently nodes, edges and labels. Also, it compares GrepVS filtering and matching phases with the state of the art graph matching algorithms.

Recent technological advances in 3D digitizing, non-invasive scanning, and interactive authoring have resulted in an explosive growth of 3D models. There is a critical need to develop new mining techniques for facilitating the indexing, retrieval, clustering, comparison, and analysis of large collections of 3D models. Shen and Makedon describe a computational framework for mining 3D objects using shape features, and addresses important shape modeling and pattern discovery issues including spherical harmonic surface representation, shape registration, and surface-based statistical inferences. The mining results localize shape changes between groups of 3D objects.

In Zhao and Yao’s opinion, while many DM models concentrate on automation and efficiency, interactive DM models focus on adaptive and effective communications between human users and computer systems. The crucial point is not how intelligent users are, or how efficient systems are, but how well these two parts can be connected, adapted, understood and trusted. Some fundamental issues including processes and forms of interactive DM, as well as complexity of interactive DM systems are discussed in this article.

Rivero, Rabuñal, Dorado, and Pazos describe an application of Evolutionary Computation (EC) tools to develop automatically Artificial Neural Networks (ANNs). It also describes how EC techniques have already
been used for this purpose. The technique described in this article allows both design and training of ANNs, applied to the solution of three well-known problems. Moreover, this tool makes the simplification of ANNs to obtain networks with a small number of neurons. Results show how this technique can produce good results in solving DM problems.

Almost all existing DM algorithms have been manually designed. As a result, in general they incorporate human biases and preconceptions in their designs. Freitas and Pappa propose an alternative approach to the design of DM algorithms, namely the automatic creation of DM algorithms by Genetic Programming – a type of Evolutionary Algorithm. This approach opens new avenues for research, providing the means to design novel DM algorithms that are less limited by human biases and preconceptions, as well as the opportunity to create automatically DM algorithms tailored to the data being mined.

Gama and Rodrigues present the new model of data gathering from continuous flows of data. What distinguishes current data sources from earlier ones are the continuous flow of data and the automatic data feeds. The authors do not just have people who are entering information into a computer. Instead, they have computers entering data into one another. Major differences are pointed out between this model and previous ones. Also, the incremental setting of learning from a continuous flow of data is introduced by the authors.

The personal name problem is the situation where the authenticity, ordering, gender, and other information cannot be determined correctly and automatically for every incoming personal name. On this paper topics as the evaluation of, and selection from five very different approaches and the empirical comparisons of multiple phonetics and string similarity techniques for the personal name problem, are remarkably addressed by Phua, Lee, and Smith-Miles.

Lo and Khoo present software specification mining, where novel and existing DM and machine learning techniques are utilized to help recover software specifications which are often poorly documented, incomplete, outdated or even missing. These mined specifications can aid software developers in understanding existing systems, reducing software costs, detecting faults and improving program dependability.

Cooper and Zito investigate the statistical properties of the databases generated by the IBM QUEST program. Motivated by the claim (also supported empirical evidence) that item occurrences in real life market basket databases follow a rather different pattern, we propose an alternative model for generating artificial data.

Software metrics-based quality estimation models include those that provide a quality-based classification of program modules and those that provide a quantitative prediction of a quality factor for the program modules. In this article, two count models, Poisson regression model (PRM) and zero-inflated Poisson (ZIP) regression model, are developed and evaluated by Gao and Khoshgoftaar from those two aspects for a full-scale industrial software system.

Software based on the Variable Precision Rough Sets model (VPRS) and incorporating resampling techniques is presented by Griffiths and Beynon as a modern DM tool. The software allows for data analysis, resulting in a classifier based on a set of ‘if ... then ...’ decision rules. It provides analysts with clear illustrative graphs depicting ‘veins’ of information within their dataset, and resampling analysis allows for the identification of the most important descriptive attributes within their data.

Program comprehension is a critical task in the software life cycle. Ioannis addresses an emerging field, namely program comprehension through DM. Many researchers consider the specific task to be one of the “hottest” ones nowadays, with large financial and research interest.

The bioinformatics example already approached in the 1e of the present volume is here addressed by Liberati in a novel way, joining two methodologies developed in different fields, namely minimum description length principle and adaptive Bayesian networks, to implement a new mining tool. The novel approach is then compared with the previous one, showing pros and cons of the two, thus inducing that a combination of the new technique together with the one proposed in the previous edition is the best approach to face the many aspects of the problem.

Integrative analysis of biological data from multiple heterogeneous sources has been employed for a short while with some success. Different DM techniques for such integrative analyses have been developed (which should not be confused with attempts at data integration). Moturu, Parsons, Zhao, and Liu summarize effectively these techniques in an intuitive framework while discussing the background and future trends for this area.
Bhatnagar and Gupta cover in chronological order, the evolution of the formal “KDD process Model”, both at the conceptual and practical level. They analyze the strengths and weaknesses of each model and provide the definitions of some of the related terms.

Cheng and Shih present an improved feature reduction method in the combinational input and feature space for Support Vector Machines (SVM). In the input space, they select a subset of input features by ranking their contributions to the decision function. In the feature space, features are ranked according to the weighted support vector in each dimension. By combining both input and feature space, Cheng and Shih develop a fast non-linear SVM without a significant loss in performance.

Im and Ras discuss data security in DM. In particular, they describe the problem of confidential data reconstruction by Chase in distributed knowledge discovery systems, and discuss protection methods.

In problems which possibly involve much feature interactions, attribute evaluation measures that estimate the quality of one feature independently of the context of other features measures are not appropriate. Robnik-Šikonja provides and overviews those measures, which are based on the Relief algorithm, taking context into account through distance between the instances.

Kretowski and Grzes present an evolutionary approach to induction of decision trees. The evolutionary inducer generates univariate, oblique and mixed trees, and in contrast to classical top-down methods, the algorithm searches for an optimal tree in a global manner. Development of specialized genetic operators allow the system exchange tree parts, generate new sub-trees, prune existing ones as well as change the node type and the tests. A flexible fitness function enables a user to control the inductive biases, and globally induced decision trees are generally simpler with at least the same accuracy as typical top-down classifiers.

Li, Ye, and Kambhamettu present a very general strategy—without assumption of image alignment—for image representation via interest pixel mining. Under the assumption of image alignment, they have intensive studies on linear discriminant analysis. One of their papers, “A two-stage linear discriminant analysis via QR-decomposition”, was awarded as a fast-breaking paper by Thomson Scientific in April 2007.

As a part of preprocessing and exploratory data analysis, visualization of the data helps to decide which kind of DM method probably leads to good results or whether outliers need to be treated. Rehm, Klawonn, and Kruse present two efficient methods of visualizing high-dimensional data on a plane using a new approach.

Yuan and Wu discuss the problem of repetitive pattern mining in multimedia data. Initially, they explain the purpose of mining repetitive patterns and give examples of repetitive patterns appearing in image/video/audio data accordingly. Finally, they discuss the challenges of mining such patterns in multimedia data, and the differences from mining traditional transaction and text data. The major components of repetitive pattern discovery are discussed, together with the state-of-the-art techniques.

Tsinaraki and Christodoulakis discuss semantic multimedia retrieval and filtering. Since the MPEG-7 is the dominant standard in multimedia content description, they focus on MPEG-7 based retrieval and filtering. Finally, the authors present the MPEG-7 Query Language (MP7QL), a powerful query language that they have developed for expressing queries on MPEG-7 descriptions, as well as an MP7QL compatible Filtering and Search Preferences (FASP) model. The data model of the MP7QL is the MPEG-7 and its output format is MPEG-7, thus guaranteeing the closure of the language. The MP7QL allows for querying every aspect of an MPEG-7 multimedia content description.

Richard presents some aspects of audio signals automatic indexing with a focus on music signals. The goal of this field is to develop techniques that permit to extract automatically high-level information from the digital raw audio to provide new means to navigate and search in large audio databases. Following a brief overview of audio indexing background, the major building blocks of a typical audio indexing system are described and illustrated with a number of studies conducted by the authors and his colleagues.

With the progress in computing, multimedia data becomes increasingly important to DW. Audio and speech processing is the key to the efficient management and mining of these data. Tan provides in-depth coverage of audio and speech DM and reviews recent advances.

Li presents how DW techniques can be used for improving the quality of association mining. It introduces two important approaches. The first approach requests users to inputs meta-rules through data cubes to describe
desired associations between data items in certain data dimensions. The second approach requests users to provide condition and decision attributes to find desired associations between data granules. The author has made significant contributions to the second approach recently. He is an Associate Editor of the *International Journal of Pattern Recognition and Artificial Intelligence* and an Associate Editor of the *IEEE Intelligent Informatics Bulletin*.

Data cube compression arises from the problem of gaining access and querying massive multidimensional datasets stored in networked data warehouses. Cuzzocrea focuses on state-of-the-art data cube compression techniques and provides a theoretical review of such proposals, by putting in evidence and criticizing the complexities of the building, storing, maintenance, and query phases.

Conceptual modeling is widely recognized to be the necessary foundation for building a database that is well-documented and fully satisfies the user requirements. Although UML and Entity/Relationship are widespread conceptual models, they do not provide specific support for multidimensional modeling. In order to let the user verify the usefulness of a conceptual modeling step in DW design, Golfarelli discusses the expressivity of the *Dimensional Fact Model*, a graphical conceptual model specifically devised for multidimensional design.

Tu introduces the novel technique of automatically tuning database systems based on feedback control loops via rigorous system modeling and controller design. He has also worked on performance analysis of peer-to-peer systems, QoS-aware query processing, and data placement in multimedia databases.

Currently researches focus on particular aspects of a DW development and none of them proposed a systematic design approach that takes into account the end-user requirements. Nabli, Feki, Ben-Abdallah, and Gargouri present a four-step DM/DW conceptual schema design approach that assists the decision maker in expressing their requirements in an intuitive format; automatically transforms the requirements into DM star schemes; automatically merges the star schemes to construct the DW schema; and maps the DW schema to the data source.

Current data warehouses include a time dimension that allows one to keep track of the evolution of measures under analysis. Nevertheless, this dimension cannot be used for indicating changes to dimension data. Malinowski and Zimányi present a conceptual model for designing temporal data warehouses based on the research in temporal databases. The model supports different temporality types, i.e., lifespan, valid time, transaction time coming from source systems, and loading time, generated in a data warehouse. This support is used for representing time-varying levels, dimensions, hierarchies, and measures.

Verykios investigates a representative cluster of research issues falling under the broader area of privacy preserving DM, which refers to the process of mining the data without impinging on the privacy of the data at hand. The specific problem targeted in here is known as association rule hiding and concerns to the process of applying certain types of modifications to the data in such a way that a certain type of knowledge (the association rules) escapes the mining.

The development of DM has the capacity of compromise privacy in ways not previously possible, an issue not only exacerbated through inaccurate data and ethical abuse but also by a lagging legal framework which struggles, at times, to catch up with technological innovation. Wahlstrom, Roddick, Sarre, Estivill-Castro and Vries explore the legal and technical issues of privacy preservation in DM.

Given large data collections of person-specific information, providers can mine data to learn patterns, models, and trends that can be used to provide personalized services. The potential benefits of DM are substantial, but the analysis of sensitive personal data creates concerns about privacy. Oliveira addresses the concerns about privacy, data security, and intellectual property rights on the collection and analysis of sensitive personal data.

With the advent of the information explosion, it becomes crucial to support intelligent personalized retrieval mechanisms for users to identify the results of a manageable size satisfying user-specific needs. To achieve this goal, it is important to model user preference and mine preferences from implicit user behaviors (e.g., user clicks). Hwang discusses recent efforts to extend mining research to preference and identify goals for the future works.

According to González Cisaro & Nigro, due to the complexity of nowadays data and the fact that information stored in current databases is not always present at necessary different levels of detail for decision-making processes, a new data type is needed. It is a *Symbolic Object*, which allows representing physics entities or real
word concepts in dual form, respecting their internal variations and structure. The Symbolic Object Warehouse permits the intentional description of most important organization concepts, follow by Symbolic Methods that work on these objects to acquire new knowledge.

Castillo, Iglesias, and Serrano present a survey on the most known systems to avoid overloading users’ mail inbox with unsolicited and illegitimate e-mails. These filtering systems are mainly relying on the analysis of the origin and links contained in e-mails. Since this information is always changing, the systems effectiveness depends on the continuous updating of verification lists.

The evolution of clearinghouses in many ways reflects the evolution of geospatial technologies themselves. The Internet, which has pushed GIS and related technology to the leading edge, has been in many ways fed by the dramatic increase in available data, tools, and applications hosted or developed through the geospatial data clearinghouse movement. Kelly, Haupt, and Baxter outline those advances and offers the reader historic insight into the future of geospatial information.

Angiulli provides an up-to-date view on distance- and density-based methods for large datasets, on subspace outlier mining approaches, and on outlier detection algorithms for processing data streams. Throughout his document different outlier mining tasks are presented, peculiarities of the various methods are pointed out, and relationships among them are addressed. In another paper, Kaur offers various non-parametric approaches used for outlier detection.

The issue of missing values in DM is discussed by Beynon, including the possible drawbacks from their presence, especially when using traditional DM techniques. The nascent CaRBS technique is expositied since it can undertake DM without the need to manage any missing values present. The benchmarked results, when DM incomplete data and data where missing values have been imputed, offers the reader the clearest demonstration of the effect on results from transforming data due to the presence of missing values.

Dorn and Hou examine the quality of association rules derived based on the well-known support-confidence framework using the Chi-squared test. The experimental results show that around 30% of the rules satisfying the minimum support and minimum confidence are in fact statistically insignificant. Integrate statistical analysis into DM techniques can make knowledge discovery more reliable.

The popular querying and data storage models still work with data that are precise. Even though there has recently been much interest in looking at problems arising in storing and retrieving data that are incompletely specified (hence imprecise), such systems have not gained widespread acceptance yet. Nambiar describes challenges involved in supporting imprecision in database systems, briefly explains solutions developed.

Among the different risks Bonafeade’s work concentrates on operational risks, which form a banking perspective, is due to processes, people, systems (Endogenous) and external events (Exogenous). Bonafeade furnishes a conceptual modeling for measurement operational risk and, statistical models applied in the banking sector but adaptable to other fields.

Friedland describes a hidden social structure that may be detectable within large datasets consisting of individuals and their employments or other affiliations. For the most part, individuals in such datasets appear to behave independently. However, sometimes there is enough information to rule out independence and to highlight coordinated behavior. Such individuals acting together are socially tied, and in one case study aimed at predicting fraud in the securities industry, the coordinated behavior was an indicator of higher-risk individuals.

Akdag and Truck focus on studies in Qualitative Reasoning, using degrees on a totally ordered scale in a many-valued logic system. Qualitative degrees are a good way to represent uncertain and imprecise knowledge to model approximate reasoning. The qualitative theory takes place between the probability theory and the possibility theory. After defining formalism by logical and arithmetical operators, they detail several aggregators using possibility theory tools such that our probability-like axiomatic system derives interesting results.

Figini presents a comparison, based on survival analysis modeling, between classical and novel DM techniques to predict rates of customer churn. He shows that the novel DM techniques lead to more robust conclusions. In particular, although the lift of the best models are substantially similar, survival analysis modeling gives more valuable information, such as a whole predicted survival function, rather than a single predicted survival probability.
Recent studies show that the method of modeling score distribution is beneficial to various applications. Doloc-Mihu presents the score distribution modeling approach and briefly surveys theoretical and empirical studies on the distribution models, followed by several of its applications.

Valle discusses, among other topics, the most important statistical techniques built to show the relationship between firm performance and its causes, and illustrates the most recent developments in this field.

Data streams arise in many industrial and scientific applications such as network monitoring and meteorology. Dasu and Weiss discuss the unique analytical challenges posed by data streams such as rate of accumulation, continuously changing distributions, and limited access to data. It describes the important classes of problems in mining data streams including data reduction and summarization; change detection; and anomaly and outlier detection. It also provides a brief overview of existing techniques that draw from numerous disciplines such as database research and statistics.

Vast amounts of data are being generated to extract implicit patterns of ambient air pollutant data. Because air pollution data are generally collected in a wide area of interest over a relatively long period, such analyses should take into account both temporal and spatial characteristics. DM techniques can help investigate the behavior of ambient air pollutants and allow us to extract implicit and potentially useful knowledge from complex air quality data. Kim, Temiyasathit, Park, and Chen present the DM processes to analyze complex behavior of ambient air pollution.

Moon, Simpson, and Kumara introduce a methodology for identifying a platform along with variant and unique modules in a product family using design knowledge extracted with an ontology and DM techniques. Fuzzy c-means clustering is used to determine initial clusters based on the similarity among functional features. The clustering result is identified as the platform and the modules by the fuzzy set theory and classification. The proposed methodology could provide designers with module-based platform and modules that can be adapted to product design during conceptual design.

Analysis of past performance of production systems is necessary in any manufacturing plan to improve manufacturing quality or throughput. However, data accumulated in manufacturing plants have unique characteristics, such as unbalanced distribution of the target attribute, and a small training set relative to the number of input features. Rokach surveys recent researches and applications in this field.

Seng and Srinivasan discuss the numerous challenges that complicate the mining of data generated by chemical processes, which are characterized for being dynamic systems equipped with hundreds or thousands of sensors that generate readings at regular intervals. The two key areas where DM techniques can facilitate knowledge extraction from plant data, namely (1) process visualization and state-identification, and (2) modeling of chemical processes for process control and supervision, are also reviewed in this article.

The telecommunications industry, because of the availability of large amounts of high quality data, is a heavy user of DM technology. Weiss discusses the DM challenges that face this industry and survey three common types of DM applications: marketing, fraud detection, and network fault isolation and prediction.

Understanding the roles of genes and their interactions is a central challenge in genome research. Ye, Janardan, and Kumar describe an efficient computational approach for automatic retrieval of images with overlapping expression patterns from a large database of expression pattern images for *Drosophila melanogaster*. The approach approximates a set of data matrices, representing expression pattern images, by a collection of matrices of low rank through the iterative, approximate solution of a suitable optimization problem. Experiments show that this approach extracts biologically meaningful features and is competitive with other techniques.

Khoury, Toussaint, Ciampi, Antoniano, Murie, and Nadon present, in the context of clustering applied to DNA microarray probes, a better alternative to classical techniques. It is based on proximity-graphs, which has the advantage of being relatively simple and of providing a clear visualization of the data, from which one can directly determine whether or not the data support the existence of clusters.

There has been no formal research about using a fuzzy Bayesian model to develop an autonomous task analysis tool. Lin and Lehto summarize a 4-year study that focuses on a Bayesian based machine learning application to help identify and predict the agents’ subtasks from the call center’s naturalistic decision making’s environment. Preliminary results indicate this approach successfully learned how to predict subtasks from the telephone con-
versations and support the conclusion that Bayesian methods can serve as a practical methodology in research area of task analysis as well as other areas of naturalistic decision making.

Financial time series are a sequence of financial data obtained in a fixed period of time. Bose, Leung, and Lau describe how financial time series data can be analyzed using the knowledge discovery in databases framework that consists of five key steps: goal identification, data preprocessing, data transformation, DM, interpretation and evaluation. The article provides an appraisal of several machine learning based techniques that are used for this purpose and identifies promising new developments in hybrid soft computing models.

Maruster and Faber focus on providing insights about patterns of behavior of a specific user group, namely farmers, during the usage of decision support systems. User’s patterns of behavior are analyzed by combining these insights with decision making theories, and previous work concerning the development of farmer groups. It provides a method of automatically analyzing the logs resulted from the usage of the decision support system by process mining. The results of their analysis support the redesigning and personalization of decision support systems in order to address specific farmer’s characteristics.

Differential proteomics studies the differences between distinct proteomes like normal versus diseased cells, diseased versus treated cells, and so on. Zhang, Orcun, Ouzzani, and Oh introduce the generic DM steps needed for differential proteomics, which include data transformation, spectrum deconvolution, protein identification, alignment, normalization, statistical significance test, pattern recognition, and molecular correlation. Protein associated data sources such as sequences, structures and interactions accumulate abundant information for DM researchers. Li, Li, Nanyang, and Zhao glimpse the DM methods for the discovery of the underlying patterns at protein interaction sites, the most dominated regions to mediate protein-protein interactions. The authors proposed the concept of binding motif pairs and emerging patterns in DM field.

The applications of DWM are everywhere: from Applications in Steel Industry (Ordieres-Meré, Castejón-Limas, and González-Marcos) to DM in Protein Identification by Tandem Mass Spectrometry (Wan); from Mining Smart Card Data from an Urban Transit Network (Agard, Morency, and Trépanier) to Data Warehouse in the Finnish Police (Juntunen)... The list of DWM applications is endless and the future DWM is promising.

Since the current knowledge explosion pushes DWM, a multidisciplinary subject, to ever-expanding new frontiers, any inclusions, omissions, and even revolutionary concepts are a necessary part of our professional life. In spite of all the efforts of our team, should you find any ambiguities or perceived inaccuracies, please contact me at j.john.wang@gmail.com.

REFERENCES


